Joint Research Programme
BTO 2023.017 | February 2023

Spectral Quality - Quality prediction of tandem mass spectra of environmentally relevant compounds using machine learning

Spectral Quality - Quality prediction of tandem mass spectra of environmentally relevant compounds using machine learning

BTO 2023.017 | February 2023

This research is part of the Joint Research Programme of KWR, the water utilities and Vewin.

Project number
402045/240

Project manager
Patrick Bäuerlein

Client
BTO - Thematical research - Chemical safety

Author(s)
Svetlana Codrean (VU), Benno Kruit (VU), Nienke Meekel, Dennis Vughs, Frederic Béen

Quality Assurance
Thomas ter Laak, Peter van Thienen

Sent to
This report is distributed to BTO-participants and is public.

Keywords

HRMS, machine learning, MS2, spectral quality

Year of publishing
2023

More information
Dr. Frederic Béen
T +31 30 606 9748
E frederic.been@kwrwater.nl

PO Box 1072
3430 BB Nieuwegein
The Netherlands

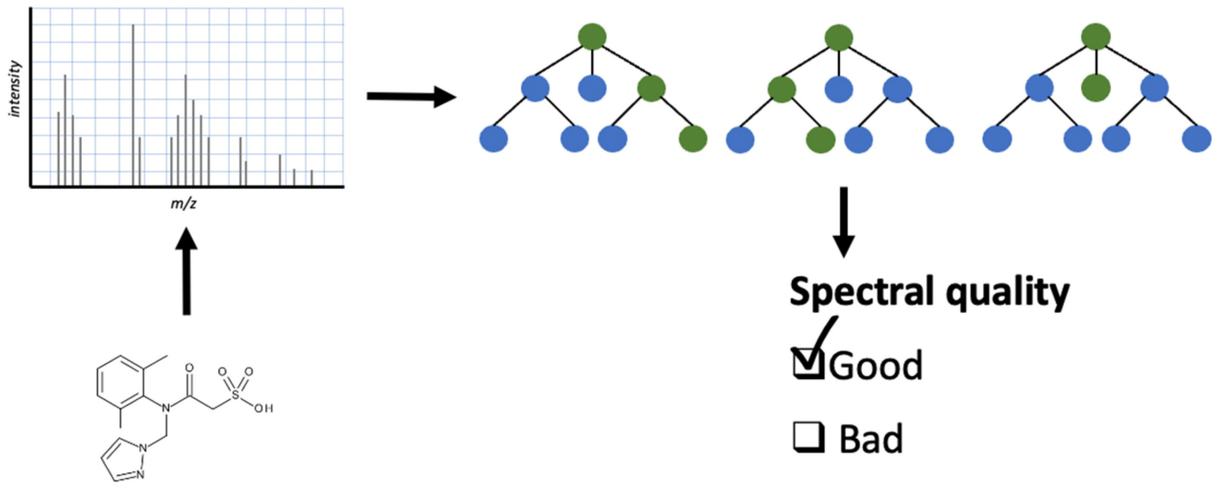
T +31 (0)30 60 69 511
E info@kwrwater.nl
I www.kwrwater.nl

Managementsamenvatting

Spectral Quality- Kwaliteit van tandem massaspectra van milieu-relevante verbindingen voorspellen met hulp van machine learning

Authors Svetlana Codrean (VU), Benno Kruit (VU), Nienke Meekel, Dennis Vughs, Frederic Béen.

De toepassingsmogelijkheden voor suspect (SS) en non-target-screening (NTS) met massaspectrometrie zijn sterk uitgebreid door de het gebruik van tandemmassaspectrometrie (MS2) in combinatie met steeds betere en volledigere bibliotheken massaspectra. Hoge resolutie massaspectrometrie (HRMS) is een vrijwel onmisbaar instrument geworden voor het monitoren van opkomende verontreinigingen in het milieu. Door een data-analyse workflow te ontwikkelen, gebaseerd op machine learning, kan de kwaliteit van tandem massaspectra nu objectief en automatisch worden beoordeeld. Het ontwikkelde algoritme kan gemakkelijk worden toegepast door drinkwaterlaboratoria om de kwaliteit van de verkregen SS- en NTS-gegevens te evalueren. Uiteindelijk zal dit leiden tot efficiëntere en minder tijdrovende gegevensanalyses en kan het aantal voorheen onbekende chemische stoffen dat wordt geïdentificeerd mogelijk toenemen.



Bijschrift: Schema van de bepaling van kwaliteit van massaspectra.

Belang: Tandem-massaspectra van goede kwaliteit zijn essentieel voor HRMS-analyses
Hoge-resolutie massaspectrometrie (HRMS) in combinatie met vloeistof- (LC) of gaschromatografie (GC) is een vrijwel onmisbaar instrument geworden voor het monitoren van opkomende verontreinigingen in het milieu. Met name de verwerving van tandemmassaspectra (MS2, waarbij verbindingen worden gefragmenteerd om informatie over hun structuur te verkrijgen) in combinatie met de steeds toenemende kwaliteit en uitgebreidheid van bibliotheken van MS2-spectra van stoffen (zoals MassBankEU, MoNA) hebben de mogelijkheden voor suspect- (SS) en non-targetscreeninganalyses (NTS) sterk uitgebreid. Toch zijn er nog veel meer potentieel relevante verontreinigingen in milieumonsters dan stoffen waarover spectrale informatie beschikbaar is in bibliotheken.

Verbetering van de kwaliteit van MS2-spectra kan zowel de annotatie van kenmerken verbeteren (wat meer succesvolle identificaties kan opleveren) als de totale verwerkingstijd van analyseresultaten verkorten. Daarnaast zijn MS2-spectra van hoge kwaliteit nodig voor toepassing van de prioriteringsstrategieën op basis van voorspellende modellering met MS2-gegevens als input. Er zijn steeds meer van dergelijke prioriteringsstrategieën en die kunnen een paradigmaverschuiving teweegbrengen van een identificatiegedreven toepassing van HRMS naar een situatie waarin men eerst informatie probeert te verkrijgen over de relevantie van een onbekende verbinding (bv. toxiciteit, polariteit, verwijderingsrendement) en pas later, indien nodig, probeert deze formeel te identificeren.

Benadering: Machine learning inzetten om de kwaliteit van tandem massaspectra te bepalen
Om een aanpak te ontwikkelen waarmee de kwaliteit van MS2-spectra automatisch en objectief kan worden beoordeeld, werd een algoritme voor machine learning gebruikt. Meer in het bijzonder werd een Random Forest (RF) classificator getraind op een set MS2-spectra van 204 referentiestandaarden van milieu-relevante verbindingen. Daarnaast werden verschillende wiskundige kenmerken uit de ruwe spectra geëxtraheerd en gebruikt als input om het model te trainen. Door deze extractie konden de uitdagingen die gepaard gaan met de heterogeniteit van MS2-spectra worden overwonnen en werden meer homogene en gestructureerde gegevens verkregen, die essentieel zijn voor algoritmen voor machine learning.

Resultaten: Model kan automatisch bepalen of een MS2-spectrum van goede kwaliteit is
Verschillende combinaties van kenmerken (wiskundige variabelen berekend uit ruwe MS2-spectra) zijn geëvalueerd om de beste prestaties te vinden. Het verkregen model werd vervolgens verder geoptimaliseerd om de hoogste precisie te verkrijgen. Daarbij werd het aanvaardbaarder (minder tijdrovend) geacht wanneer een MS2-spectrum van goede kwaliteit ten onrechte als slecht werd bestempeld dan omgekeerd. Het ontwikkelde algoritme streeft ernaar de verwerkingstijd te minimaliseren, zodat het investeren van tijd en middelen om een verbinding te identificeren waarvan de MS2-gegevens van slechte kwaliteit zijn moet worden vermeden. Het geoptimaliseerde model was in staat 85% van de goede spectra als zodanig correct te labelen, zodat slechts 15% van de spectra van goede kwaliteit als slecht werden bestempeld. Deze resultaten zijn in overeenstemming met andere algoritmen die zijn ontwikkeld op het gebied van proteomics, wat de robuustheid van de ontwikkelde aanpak verder onderschrijft.

Implementatie: Aanpak die verwerking suspect- en non-targetscreeningdata vergemakkelijkt
Vanuit het perspectief van een drinkwaterlaboratorium kan het binnen dit project ontwikkelde model gemakkelijk worden toegepast om de analyse en interpretatie van zowel suspect screening (ook vaak bibliotheekscreening genoemd) als NTS-gegevens te vergemakkelijken. Het ontwikkelde algoritme moet met name de nabewerking van gegevens vergemakkelijken door een betere prioritering waarbij alleen aandacht gaat naar kenmerken met MS2-gegevens van goede kwaliteit. Ook kan het objectieve informatie verschaffen om te beslissen of specifieke monsters opnieuw moeten worden geanalyseerd (bijvoorbeeld wanneer kenmerken die prioriteit hebben gekregen MS2-data van lage kwaliteit blijken te hebben) en moet het de kansen op succesvolle (voorlopige) identificatie via bibliotheekonderzoeken vergroten. Om de ontwikkelde aanpak te gebruiken hoeven laboratoria alleen de MS2-gegevens die ze tijdens hun analyses hebben verkregen te exporteren en het ontwikkelde script te gebruiken om het model te trainen en hun gegevens te evalueren.

Rapport
Van dit onderzoek wordt verslag gedaan in het rapport *Spectral Quality - Quality prediction of tandem mass spectra of environmentally relevant compounds using machine learning* (BTO2023.017).

.

Contents

1 Introduction

High-resolution mass spectrometry (HRMS) coupled with either liquid (LC) or gas chromatography (GC) has become an almost essential tool to monitor emerging contaminants in the environment (Hollender et al., 2017). In particular, the acquisition of tandem mass spectra (MS2) combined with the ever growing quality and comprehensiveness of spectral libraries (e.g., MassBankEU (Neumann et al., 2022), MoNA (Fiehnlab, 2022)) have greatly expanded the possibilities offered by suspect and nontarget screening analyses (Mohammed Taha et al., 2022; Oberacher et al., 2019). Despite continuous improvements, large discrepancies still exist between the number of potentially relevant contaminants present in environmental samples and those for which spectral information is available in libraries (Oberacher et al., 2020). Moreover, despite the development of workflows to automatically improve the quality of records added to these libraries (Stravs et al., 2013) and the acquisition of multiple spectra per compound to account for specific fragmentation curves, some issues regarding quality assurance and control (QA/QC) of the information contained in these databases still exist, including insufficiently curated tandem mass spectra (Oberacher et al., 2020; Schulze et al., 2020). In the field of proteomics, where database searches and *de novo* sequencing approaches are used to identify peptides from complex mixtures of proteins (Gholamizoj and Ma, 2022; Ma, 2017; Nesvizhskii et al., 2006),quality of tandem mass spectra and how to assess it has been the subject of various researches. Spectra of good quality consist of spectra which contain diagnostic information about the fragmented parent ion. Specifically, they should have enough peaks (i.e., *m/z* values) spread across the whole mass range (relative to the mass of the parent ion) and with sufficient intensity, as well as little to no noise. In fact, for for library-based peptide identifications, poor MS2 data quality is considered to play a major role in the occurrence of false negatives (Gholamizoj and Ma, 2022). For this purpose, already in the early 2000s, algorithms have been devised to try to automatically assess the quality of MS2 spectra acquired in proteomics experiments (Bern et al., 2004). Recently, more advanced machine and even deep learning algorithms have been developed to automatically assess the quality of acquired MS2 signals, reduce the occurrence of false negatives and decrease overall processing time of large datasets (Ding et al., 2009; Gholamizoj and Ma, 2022; Zou et al., 2009). The proposed classifiers showed very promising results. For instance, the approach developed by Bern et al (Bern et al., 2004) was able to eliminate over 75% of spectra considered as being of bad quality and, at the same time, would only lose 10% of spectra deemed as being of good quality. Using more advanced machine learning algorithms (e.g., support vector machine (SVM) and k-means), Zou et al. (Zou et al., 2009) and Ding et al. (Ding et al., 2009) were able to develop binary classifiers having true positive rates (TPR) of 92.1% and 90% while keeping the true negative rate (TNR) at 89.6% and 92%, respectively. These methods often relied on a range of “features” (to be understood here as descriptors, or independent variables, rather than HRMS-based features) derived from peptide fragmentation patterns, such as b- and y-ion peaks (Choo and Tham, 2007) or amino acid sequence tags (Nesvizhskii et al., 2006).

Only more recently, a deep learning method was developed which takes the entire MS2 spectrum (after pre-processing and normalisation) to assess spectral quality (Gholamizoj and Ma, 2022). The fact that most models developed so far used features derived from specific peptide fragmentation patterns, combined with the difficulty to objectively establish criteria to define an MS2 spectrum of good quality, might explain why these approaches have not been implemented in other fields. In fact, in the specific case of (small) environmentally relevant molecules, the issue of MS2 spectral quality has not been addressed thoroughly, besides in the general context of curating spectral libraries and the development of search and matching algorithms (Oberacher et al., 2020). Yet, obtaining MS2 spectra of good quality would both improve feature annotation and reduce overall (post-)processing time in environmental analyses. However, the importance of obtaining high quality MS2 spectra is not limited to annotations and library searches. In fact, in recent years, an increasing number of computational tools have been reported which make use of MS2 data to improve post-processing and prioritisation (e.g., molecular networking strategies (Oberleitner et al., 2021; Watrous et al., 2012)), predict molecular structures (Dührkop et al., 2015) or

even *in vivo* toxicity of unknowns (Peets et al., 2022). Given that these methods rely on MS2 spectra, their performances would greatly benefit from having input data of high(er) quality. Furthermore, algorithms to automatically assess MS2 quality could in future be integrated in HRMS acquisition methods as additional criteria to trigger further fragmentation of ions selected during MS1 survey scans in data-depended acquisition (DDA) mode. More specifically, if an ion is for instance isolated and fragmented because its intensity is above a certain threshold, but the obtained MS2 spectrum is classified as of insufficient quality by the algorithm, then an additional fragmentation (e.g., with a different CE) can be triggered. Whilst for data independent acquisition (DIA), such information could be useful during post-processing to prioritise MS2 spectra rich in information.
Building on the promising results obtained in the field of proteomics and the added value that automated prediction of MS2 quality would have in the field of small molecules, this work focused on the development of a machine learning pipeline to automatically assess the quality of mass spectra of environmentally relevant compounds. For this purpose, a dataset of 204 reference standards of environmental contaminants acquired with different collision energies (CEs), corresponding to almost 1400 MS2 spectra, was used. Initially, focus was set on finding relevant and non-redundant descriptors which could be used for machine learning purposes and that provided a sufficiently accurate representation of the raw input data. Specifically, three different feature sets were computed, and their performances were evaluated using a Random Forest (RF) Classifier with cross-validation. Computed descriptors were then further filtered to select those which explained most of the available data. Finally, the optimised feature sets were evaluated against the test set and a final classification model was optimised to discriminate between MS2 spectra of good and bad quality.

From a drinking water quality perspective, the ultimate goal of this project consisted in developing an algorithm that can be readily implemented by drinking water laboratories to evaluate the quality of the acquired suspect (SS) and non-target screening (NTS) data. In particular, the developed algorithm is supposed to facilitate data post-processing (i.e., after actual acquisition) through an improved prioritisation (i.e., focus only on those features which have MS2 data of good quality), provide objective information to decide whether to reanalyse specific samples (i.e., should features that have been prioritised for one or another reason have MS2 data of poor quality) and, last but not least, increase chances of successful (tentative) identification via library searches. The latter would eventually reduce time spent on trying to identify features as well as reduce cost associated with purchasing reference material (as one would ideally focus on features with a high chance of being successfully identified).

2 Materials and methods

2.1 Dataset

The dataset used in this work consisted of fragmentation mass spectra (MS2) of 204 reference standards of known environmental contaminants (obtained in the BTO project 402045-151 Non-target screening op tijd en kwantitatief) which were analysed by liquid chromatography (LC) coupled to an Orbitrap Fusion Tribrid high-resolution mass spectrometer (HRMS, Thermo Fisher Scientific). Separation was achieved using a generic chromatographic method using an XBridge BEH C18 (2.5 µm, 2.1 × 100 mm Column XP, Waters) column as described in Been et al. (2021). Acquisition was performed in data-dependent acquisition (DDA) mode with high collision dissociation (HCD) and graded collision energy (CE) of 10, 20, 35, 50, 65, 80 and 100%. MS2 spectra obtained were then searched using the retention time of each reference standard and by retrieving the scan corresponding to each of the CEs used. Spectra were acquired in profile mode but where then converted to centroids to facilitate comparison with existing spectra libraries. The final dataset consisted of 1399 MS2 spectra.

2.2 Initial labelling of MS2 spectra

Initially, labelling of acquires MS2 spectra was carried out automatically. More specifically, matching spectra were searched in MassBankEU (Schulze et al., 2021) using the *SpectrumSimilarity* function from the OrgMassSpecR package developed by Dodder and Mullen (2017). Spectra eliciting a high score (≥ 0.75) were initially labelled as being of "GOOD" quality while spectra with lower quality were labelled as "BAD". However, due to the differences in both fragmentation approaches and collision energies (CEs) used, inconsistencies were observed in the labelling. In particular, spectra were incorrectly labelled. Because of the difficulty of defining quantitative criteria which could be used to automatically label MS2 spectra, it was decided to rely on expert judgement and to manually label all spectra. In particular, the number of fragments, their distribution across the *m/z* range (with respect to the mass of the molecular ion) and intensity with respect to the molecular or base ion were used as criteria to define whether a spectrum could be considered of GOOD or BAD quality.

2.3 Pre-processing

Prior to calculating features (i.e., descriptors), MS2 spectra were scaled both with respect to their intensity and *m/z* range. Specifically, relative intensities (i.e., range [0,1]) were computed by dividing individual intensities by the intensity of the base peak (i.e., the most intense peak in the MS2 spectrum). Similarly, the *m/z* range of each spectrum was normalised by dividing individual *m/z* values by the m/z value of the precursor. Finally, noise was removed by filtering all *m/z* ratios whose intensity was ≤ 5% of the base peak. An overview of the distribution of the pre-processed MS2 spectra is shown in Figure 1. Results show that the distribution of relative *m/z* values in BAD spectra appear to be slightly more skewed towards lower values compared to spectra labelled as GOOD.

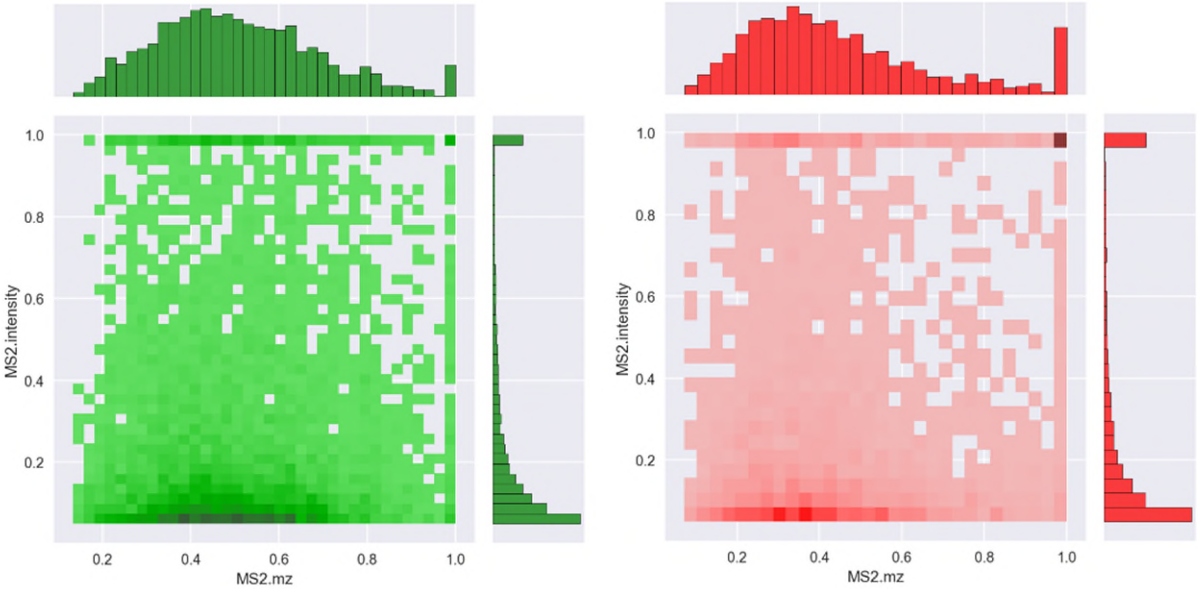


Figure 1: Distribution of m/z and intensities in MS2 spectra labelled as GOOD (left) and BAD (right).

2.4 Feature transformation

To develop a machine learning algorithm that can efficiently classify tandem mass spectra based on their quality, three different set of *features* were computed from pre-processed spectra. It should be noted that in the context of this work, the term *feature* is used to refer to mathematical variables computed or extracted from MS2 spectra and not *features* as used in the context of suspect and non-target screening (i.e., accurate mass, retention time and peak intensity).

2.4.1 Distance features

The first set of features which were computed consisted of statistics derived from the calculation of Euclidean distance between the centroid of each spectrum and the remaining *m/z* after pre-processing. The centroid *c* was defined as follows

$$c = (\frac{\sum_{j=1}^{n} m_j}{n}, \frac{\sum_{j=1}^{n} i_j}{n}) \qquad \text{eq.1}$$

where m_j are the *m/z* values in the spectrum, i_j are the corresponding intensities and *n* is the number of *m/z* values in the spectrum. For every *m/z* value (*p*) in the spectrum, the Euclidean distance *d* to the centroid *c* is calculated by the formula

$$d = \sqrt{(m_c - m_p)^2 + (i_c - i_p)^2} \quad \text{eq.2}$$

where m_p and i_p are the *m/z* and corresponding intensity of the p^{th} *m/z* value in the spectrum. Using the distance vector, the count, mean, median, standard deviation, minimum, maximum, first, second and third quartiles were calculated and used as *distance features* for data processing.

2.4.2 Handcrafted features

The second set of features computed from MS2 spectra consists of a collection of the most common features found in the literature together with some empirically selected features. Among these is the number of *m/z* values in the spectrum which has been commonly used in previous works (Bern et al., 2004; Nesvizhskii et al., 2006; Tabb et al., 2001; Zou et al., 2009). Furthermore, the average (Nesvizhskii et al., 2006), sum (Tabb et al., 2001) and standard deviation of intensities in each spectrum were computed. Additionally, the dot product between *m/z* and intensity values was computed and used as feature. The number of peaks with relative intensity greater than 0.1 (Zou et al., 2009) and 0.2 were also considered. Two other features used were the standard deviation of the consecutive mass gaps between all peaks and the average number of peaks in a 2 Dalton (Da) interval (Nesvizhskii et al., 2006). The intensity balance was calculated by dividing the *m/z* axis into a number of bins of equal width and subtracting the total intensity of the first bin from the sum of the intensities of the remaining bins (Bern et al., 2004). Finally, the Shannon entropies for the *m/z* vector and the intensity vector were calculated.

2.4.3 Grid features

The last set of features used was inspired by previous work from Logan et al. (2004) and consisted in dividing the spectra in 1 or 2 dimensional (1D or 2D) grids and counting the number of points (i.e., m/*z*) in each grid cell. In 1D grids, between 1 and 20 bins were unevenly distributed along the intensity (y) axis to have more granularity (i.e., more frequent bins) at lower intensities compared to higher ones. In the case of 2D grids, between 1 and 20 bins were considered both for the *m/z* (x) and the intensity (y) axis (i.e., yielding *N x N* matrices*)*.

2.5 Statistical modelling

Given a transformed MS2 spectrum X_j = $F_i(x_j)$, according to the feature extraction method F_i (as described above), the goal consisted in modelling the target $y_i \in \{0,1\}$, namely a quasi-Bernoulli random variable, representing the quality of an MS2 spectrum (referred to as BAD or GOOD, respectively) conditionally dependent on *X*. To predict *y* it was hence necessary to estimate $P(y|X)$. To simplify the statistical assumptions imposed by Bernoulli-like random variables and common statistical learning methods, we assumed independently, identically distributed samples, which is reasonable given the previously described data acquisition process. In the context of this work, a RF algorithm (Liu et al., 2012) was used to estimate $P(y|X)$ given its widespread use in the context of binary classification (Ali et al., 2012; Ham et al., 2005) and the fact that it is often considered the method of choice with expected highly non-linear relationships. The RF algorithm works by first bootstrapping the dataset and fitting individual classification trees to the bootstraps. While individual trees are weak learners due to their high variance, combing their predictions (i.e., majority vote) yields a substantially stronger performance (ensemble learning). When compared to more classical methods like Logistic Regression or Naive Bayes, it typically performs well "offthe-shelve" (i.e., its hyperparameters do not need to be adjusted exhaustively to work reasonably). This is a favourable property of the random forest and the reason it was chosen as main model estimator as the focus is on evaluating the predictive power of different feature sets. Also note that the RF is in fact well suitable for parameterization of $\mathbb{P}\,(y)$ since in the context of binary classification, a random forest can be used as a pseudo-density-estimator for which one can simply use the ratio of trees with positive predictions to negative predictions (i.e., the probability of success). Data processing, analysis and machine learning were performed in Python and the code can be found on GitHub (Codrean, 2022).

2.6 Validation and testing

The initial dataset of 1399 MS2 spectra was divided into two parts, namely 949 (67.8%) observations for training and 450 (32.2%) observations were kept for final testing. Both sets had 44% instances labelled (based on expert judgement) as GOOD and 56% labelled BAD, as described in 2.2. Feature groups were evaluated individually and then results were compared. Stratified 10-Fold Cross-Validation (Purushotham and Tripathy, 2012) was used over the 949 instances provided for training purposes, ensuring that the same class proportions were maintained at each split. This resulted in 854 instances for fitting and 95 samples for prediction. Metrics used to evaluate model performances were the iteration accuracy, average precision, Area Under the Receiver Operating Characteristic Curve (ROC AUC) score and the log loss of the model. The accuracy indicates the proportion of correctly identified samples. A more important metric is precision, which is the ratio of correctly classified high-quality spectra to all spectra classified as GOOD. In this case, the cost of classifying a BAD spectrum as GOOD is higher than the other way around. This type of misclassification leads to useless further analysis of an inferior spectrum that is considered of high quality, and time is wasted trying to find a match of the spectrum in the MS databases. Similarly, should this kind of classification algorithm be implemented during acquisition (i.e., determining whether an additional MS2 spectrum needs to be acquired in a DDA experiment), a conservative approach would involve the collection of an additional MS2 spectrum, despite it being already of sufficient quality, rather than having only a spectrum of low quality. For this reason, focus was set on improving precision. After evaluating the different models (i.e., different feature sets, their combination and feature-specific configurations), the best performing candidate feature sets were tested on the holdout (test) set containing the remaining 450 MS2 spectra (not used for training and validation). This is necessary to avoid "over-tuning" during cross-validation and allow a fair comparison between methods and ultimately get a realistic impression of the capabilities of the proposed method(s).

3 Results and discussion

3.1 Validation results

3.1.1 Grid features evaluation

Prior to evaluating model performances on the holdout (test) set, the optimal number of bins in both the 1D (unevenly distributed) and the 2D grids were evaluated. First, the optimal grid specification was searched, namely the number of bins per axis (*m/z* and intensity) from which the 2D distribution of the *m/z*-intensity pairs is obtained. From this distribution, specified by the number of bins on each axis, *N x N* features were derived, as described previously. Combinations of *m/z* and intensity bins from 1 to 20 were evaluated. It is worth mentioning that the pair (1, 1) means that there is only one bin for the m/z values and one bin for the intensity values and hence corresponds to the number of fragments in a spectrum. The heatmaps in Figure 3 show the results for all metrics. A first observation is that the use of a very granular grid seems to make little sense, as the heat map shows undesirable performance scores for more granular binning (Figure 2 and Figure 3). A closer look reveals an almost identical pattern in all four metrics. Areas with highest scores (i.e., darkest shades) are in two locations in the 2D space. In the case of log loss, it is the opposite, as one seeks to obtain the smallest metric. Visual inspection indicates that the best-performing pairs are (*m/z*, 1), ∀x ∈ {8, 9, …, 20}, but also the pairs (*m/z*, *intensity*), *m/z* ∈ {1,2}, *intensity* ∈ {11,12}. These results suggest that the use of 1D histograms is preferable to 2D histograms. One possible explanation could be that the more granular the space becomes, the sparser the grid cells are (i.e., most values are equal to zero). The four best bin combinations (see Table 1) were selected for further comparisons. Results obtained using the 1D unevenly distributed grid are shown in Figure 2. In this specific case, no difference was observed as the number of bins was increased up to 20, suggesting that the granularity of the lowest layers does not play an important role, likely because noise (i.e., m/z values having an intensity < 5% of the maximum) was removed during pre-processing. Nevertheless, the best performing bin dimension was 14 (i.e., accuracy of 0.71 (0.04), average precision of 0.68 (0.04), ROC AUC of 0.77 (0.04) and log loss of 9.93 (1.41)).

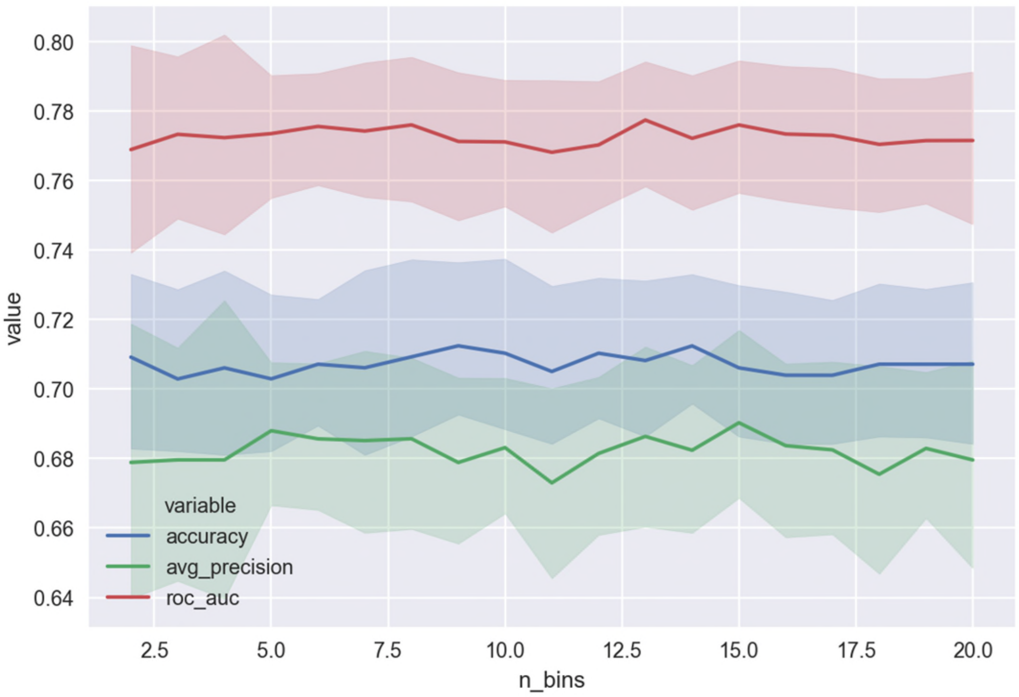


Figure 2: Results obtained using the 1D uneven grid.

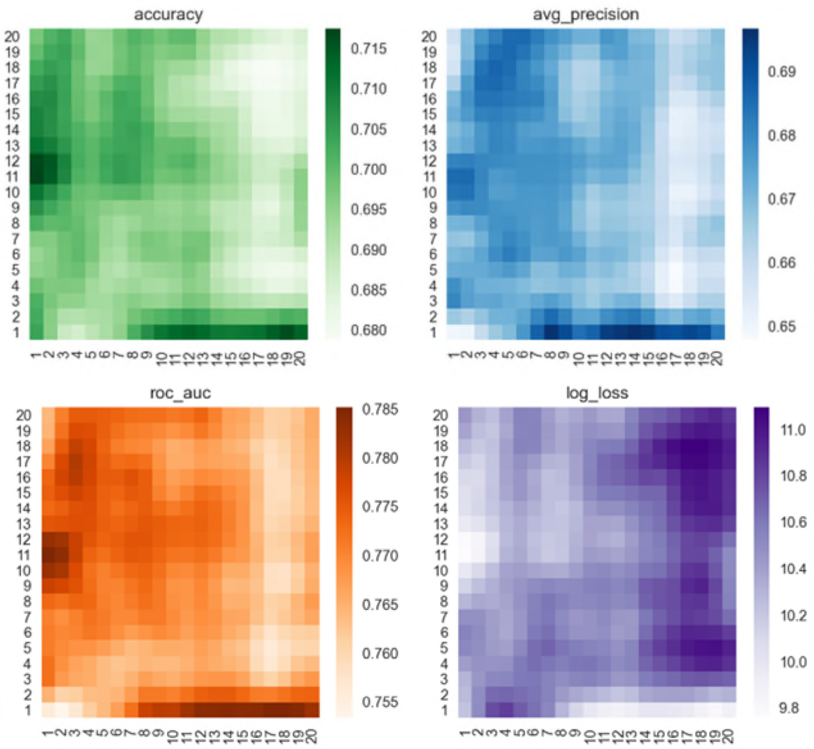


Figure 3: Metrics for each combination of m/z (x-axis) and intensity (y-axis) bins. Each value represents the average metric score (together with the standard deviation) obtained from a Stratified 10-Fold Cross-Validation for a given (#m/z, #I) combination. A Gaussian blur filter (σ = 1) was applied to the heatmap to facilitate visualisation of the results.

Table 1: Bin number combinations providing the best metrics. Results are sorted by log loss and decreasingly by average precision, ROC AUC and accuracy. Standard deviations are shown between brackets.

(# m/z bins, # intensity bins)	Accuracy		Average Precision		ROC AUC		Log Loss	
(19, 1)	0.75	(0.05)	0.73	(0.06)	0.81	(0.04)	8.77	(1.56)
(10, 1)	0.74	(0.03)	0.69	(0.06)	0.79	(0.04)	9.06	(1.03)
(12, 1)	0.73	(0.05)	0.74	(0.06)	0.80	(0.05)	9.28	(1.65)
(2, 11)	0.73	(0.04)	0.70	(0.09)	0.79	(0.06)	9.28	(1.49)

3.1.2 Features selection

In addition to the previously described *Grid* features, *Handcrafted* and newly proposed *Distance* features were tested individually and combined, as shown in Table 2, before evaluating them with the holdout test set. It should be noted that grid features were not included in the correlation testing because their structure is inherently different, while both *Distance* and *Handcrafted* features are based on heuristics and are likely going to contain similar information because criteria for handcrafting were partially similar. A Spearman rank correlation test was applied to the combined features and results are shown in Figure 4. As expected, the number of peaks and the count of distances are fully correlated. It can be observed that almost all distance features form together the cluster to the right. Interestingly, the two least correlated features are the precursor *m/z* and the collision energy. In fact, one might have expected the two to be somehow correlated as larger molecules would need higher collision energies to obtain satisfactory MS2 spectra.

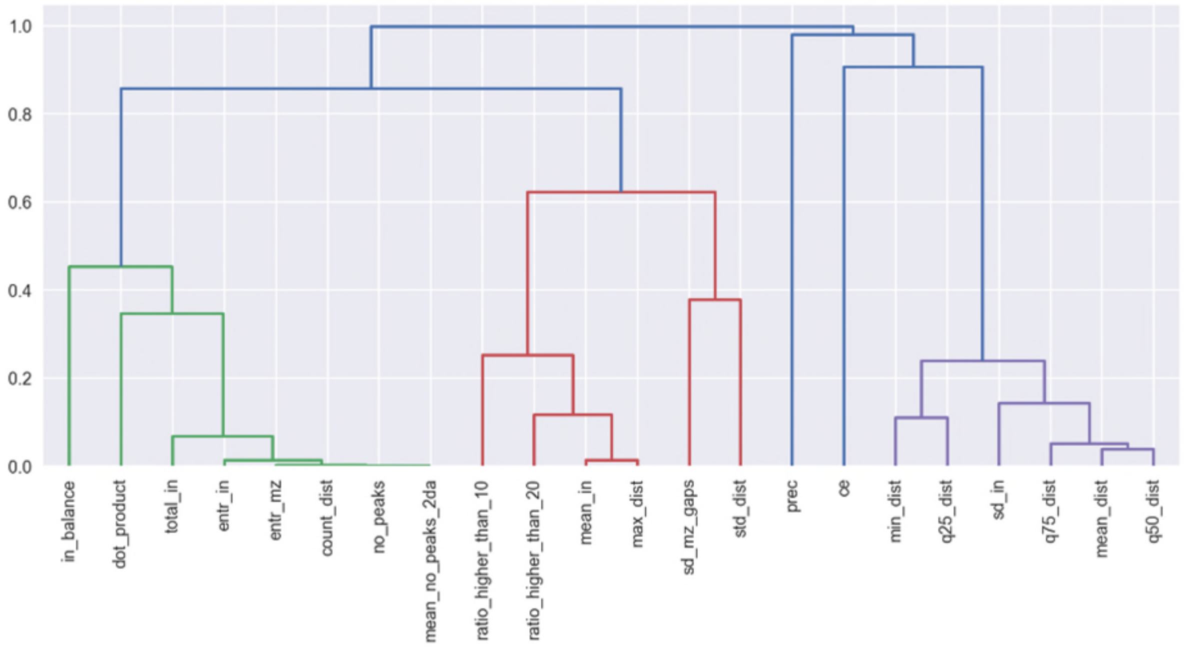


Figure 4: Hierarchical clustering dendrogram based on outcomes of the Spearman correlation test. The y-axis represents the degree of dissimilarity between the features, which is D = 1 − |ρ|, where ρ is the pairwise rank correlation coefficient. Consequently, if the correlation is one or minus one (i.e. fully correlated), the dissimilarity is zero.

Cross Validated Recursive Feature Elimination (RFECV) (Chen and Jeong, 2007) was used as a second approach for feature selection. This routine first trains a model with all features, giving each feature a rank based on its importance calculated by the estimator, in our case the Random Forest, so that features with low rank are recursively discarded until only equally important features (i.e., rank 1) remain. While this method generally depends heavily on the model's estimate of feature importance and is generally not safe as a feature selector alone, it is useful for creating new subsets of features that are then evaluated in a separate procedure. In this case, ten different feature groups were selected for evaluation, in which the combination of RFECV with a correlation removal (applied together or separately) was tested. The validation results (in the form of the Stratified 10-Fold CV) are sorted by Log Loss and presented in Table 2. The *Combined* set refers to the *Handcrafted* and *Distance* features taken together. Precursor mass and collision energy were added to all 10 chosen sets. The *2D-* and *1D-Grid* feature sets were computed using the number of bins that give the highest performances (see previous subsection). *Uncorrelated* refers to feature subsets which passed through the correlation analysis (i.e., features with a dissimilarity value below 0.3 were discarded). Recursive feature elimination (RFE) was applied only to the *Combined* set, given that none of the features were discarded for the other sets. As can be seen, best performances were obtained with the *Handcrafted* feature set, consisting of 14 features in total.

Table 2: Validation performance results of combinations of feature sets. Average and standard deviation (between brackets) are shown.

Feature groups	Accuracy		Avg Precision		ROC AUC		Log Loss		#Feats
Handcrafted	82.6	(2.3)	84.5	(4.1)	89.5	(2.7)	6.00	(0.78)	14
Combined + RFE	82.1	(3.4)	83.8	(3.9)	88.6	(3.0)	6.18	(1.16)	16
Combined	81.7	(3.9)	84.1	(4.1)	88.8	(2.9)	6.33	(1.34)	22
Handcrafted + Uncorrelated	81.6	(3.2)	84.2	(3.8)	89.0	(2.8)	6.36	(1.10)	8
Combined + Uncorrelated	81.4	(3.3)	84.5	(3.1)	89.1	(2.5)	6.44	(1.13)	9
Combined + Uncorrelated + RFE	79.8	(3.6)	82.7	(3.9)	87.5	(2.8)	6.98	(1.22)	7
Distance	78.6	(3.4)	83.4	(4.8)	87.7	(4.1)	7.38	(1.19)	10
Distance + Uncorrelated	78.5	(4.6)	80.9	(6.1)	86.9	(3.9)	7.42	(1.57)	5

1D Grid (14)	78.5	(4.6)	81.3	(4.5)	86.6	(3.1)	7.42	(1.58)	16
2D Grid (19, 1)	77.2	(3.5)	80.3	(3.0)	85.6	(2.4)	7.86	(1.22)	21

3.2 Testing results

The most promising sets of features from each category (*Handcrafted, Distance* and their combination*, 1D* and *2D-Grid*) based on validation results, were compared using the holdout (test) set. For this purpose, all models were re-trained using both training and validation sets. Results are reported in Table 3 and Figure 5. As can be seen, all feature sets perform reasonably well, achieving on average an accuracy of about 79%, an average precision of 82% and a ROC AUC of about 86%. It is interesting to notice that the baseline (i.e., number of peaks) only has an overall 10% lower performance compared to the other features. This might suggest that the number of peaks in MS2 spectra after normalisation and noise removal is a rather good predictor of spectral quality. Regarding newly introduced *Distance* and *Grid* features, these showed similar results to the *Handcrafted* features derived from previous studies. These findings are also visible in Figure 5, which shows both ROC and Precision-Recall curves. Unlike the baseline approach, the selected feature sets provided similar performances, especially with regard to the ROC curve. It is noteworthy to mention that even though obtaining relevant features in the field of small molecules is more complex compared to proteomics, where one can rely on additional information/patterns due to the occurrence of repeating units (i.e., amino acids and peptides), results obtained here are consistent with performances reported in the literature. For instance, in the recent approach proposed by Gholamizoj and Ma (2022), ROC AUC ranging from 68% to 89% were obtained for the classification of MS2 spectra of peptides.

Table 3: Performances (in %) of the models trained using the selected feature sets.

Features	Accuracy	Average precision	ROC AUC	Number of Features
Handcrafted	80.4	85.8	88.2	14
Combined + RFE	79.1	82.6	87.4	16
Distance	78.2	81.5	86.1	10
2D Grid (19, 1)	78.2	81.1	86.1	21
1D Grid (14)	78.9	81.1	86.0	16
Baseline (# peaks)	69.8	64.4	76.5	1

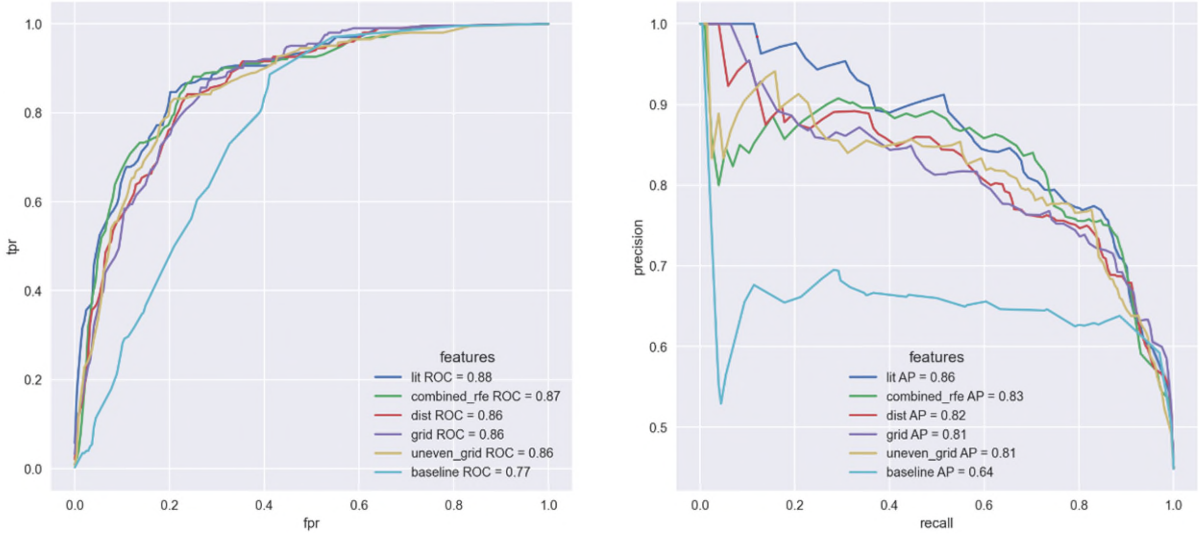
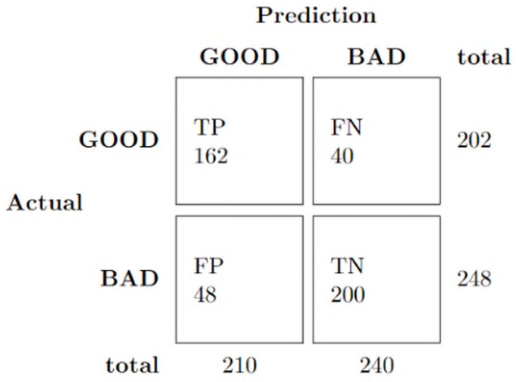


Figure 5: ROC curves (left) and Precision-Recall curves (right) obtained for the models trained using the selected feature sets.

3.3 Optimized model

Based on the outcomes of the testing step, the model with the best performances, namely the one computed using the *Handcrafted* feature set, was further investigated, and optimized. For this purpose, the confusion matrix of the Random Forest Classifier was directly examined instead of assessing the metrics derived from it. Figure 6 depicts the confusion matrix, which reveals a precision of $\frac{TP}{TP+FP} = {}^{162}/_{210} = 0.771$ using a standard threshold of 0.5. However, as discussed previously, it was decided to favour precision above the other performance parameters to minimize the chance of having a BAD spectrum being mislabelled as GOOD. For this purpose, the *f-beta* score (Goutte and Gaussier, 2005) was evaluated to find an optimal threshold for probability predictions. The *f-beta* score (i.e., $F_\beta = (1+\beta^2) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall}$) uses the beta parameter as weight for the recall. Therefore, when beta is less than one, recall is less important and the precision is more important in the f-score, which is the harmonic mean between these two scores. By moving the threshold to the right, we can create a more conservative decision maker, i.e. will assign less GOOD and more BAD. Using a *beta* parameter of 0.5, corresponding to a threshold of 0.661 and allowing to get a final accuracy of 79%, a higher precision of 85% could be attained with a recall of 65%, which is in line with other quality prediction models developed in the field of proteomics. These results are also satisfying when considering the purpose for which this model was developed, namely to automatically evaluate the quality of MS2-spectra of (small) environmentally relevant compounds. Given that laboratories using NTS analyses detect large numbers of unknown features, of which only a small fraction can generally be identified due to, among other things, low quality MS2 data, having an algorithm that can automatically label features having good MS2 data with 85% precision is highly valuable. This could in fact be used to prioritise features (i.e., focusing only on those which have higher chances of being successfully identified) and/or it could be used to decide whether additional analyses focusing on specific features (e.g., targeted experiments or inclusion lists) are required to obtain data of higher quality.



		Prediction		
		GOOD	BAD	total
Actual	GOOD	TP 162	FN 40	202
	BAD	FP 48	TN 200	248
	total	210	240	

Figure 6: Confusion matrix of the RF Classifier computed using Handcrafted features. Here the classification threshold was set to default 0.5.

4 Conclusion and recommendations

Acquisition and processing of high-quality tandem mass spectra has clear advantages, both for identification and predictive modelling purposes. However, while various applications have been reported in the field of proteomics, an automated approach to assess the quality of fragmentation spectra in the field of small and environmentally relevant molecules was still missing. In the context of this work, a RF classifier was trained capable of attaining comparable if not superior performances compared to approaches previously reported in the field of proteomics. In fact, best performing model obtained in this work provided very similar results compared to the deep learning model recently developed by Gholamizoj and Ma (2022) (0.88 and 0.89 ROC AUC). Similarly to the work done by Nesvizhskii et al. (2006), the classifier was not affected by the presence of potentially correlated features. With respect to results obtained using the *Grid* features-based model, the Random Forest classifier obtained here using a 1D grid outperformed the Gaussian Mixed model developed by Logan et al. (2004) (0.86 and 0.76 ROC AUC, respectively). Similarly, the model developed here also performed slightly better compared to the one obtained through Boosting when using a 2D grid (0.86 and 0.85 ROC AUC, respectively). Despite being developed on what might be considered a rather small dataset, obtained results suggest that the sets of features tested in the context of this work and the optimised classifier can become a very useful tool to automatically assess the quality of MS2 spectra of small and environmentally relevant molecules. Applications could range from improving and automating spectral library curation and identification, MS-based predictive modelling and even acquisition, should these approaches become part of acquisition parameters in DDA-methods for instance. Future work should focus on evaluating the performances of the obtained model on a larger dataset or use the current model in a semi-supervised approach to label a larger dataset and eventually train a more advanced model (e.g., deep learning).

From the perspective of a drinking water laboratory, the model that has been developed in the context of this project can be readily implemented to facilitate the analysis and interpretation of both suspect (often referred to as library screening) and non-target screening data. Firstly, data-dependent acquisition (DDA) data that has been acquired by laboratories for their yearly screening analysis of drinking water sources can be evaluated using the developed algorithm. Specifically, MS2 spectra of detected features can be automatically labelled as being of high or low quality using this model. The analyst can then focus only on those features which were scored as of high quality and either disregard the other ones or decide to rerun certain samples to obtain MS2 spectra of better quality (for instance through an inclusion list which would guarantee that one of multiple MS2 spectra of features of interest are acquired). This would reduce time spent on trying to identify chemical features of low quality. Similarly, it is expected that this algorithm will increase the fraction of investigated features which give a higher similarity score with libraries used (*in-house*, vendor and open access ones). Similarly, laboratories can use the developed algorithm to screen the information included in libraries used in their workflows to determine if the data they compared their analysis results to is of adequate quality or not. Secondly, given that also within the practice of environmental laboratories, including in the water sector, predictive tools based on MS2 data there is a tendency to increasingly use MS2 data for predictive purposes (e.g., toxicity, semi-quantification), having an approach which objectively and automatically allows to determine the quality of the available data is potentially of high value. In fact, prioritisation of features of interest based on these predictive models will greatly benefit from having input data of high quality. Based on the promising results obtained in this project, we encourage laboratories to test the developed algorithm on their data and to evaluate to which extent it reduces NTS processing time and whether it has an impact on the number of features that can be (tentatively) identified. The source code to process MS2 data, calculate features and train the model can be shared with the laboratories.

5 References

Ali, J., Khan, R., Ahmad, N., Maqsood, I., 2012. Random Forests and Decision Trees. Int. J. Comput. Sci. Issues 9, 272.

Been, F., Kruve, A., Vughs, D., Meekel, N., Reus, A., Zwartsen, A., Wessel, A., Fischer, A., ter Laak, T., Brunner, A.M., 2021. Risk-based prioritization of suspects detected in riverine water using complementary chromatographic techniques. Water Res. 204, 117612. https://doi.org/10.1016/j.watres.2021.117612

Bern, M., Goldberg, D., McDonald, W.H., Yates, J.R., 2004. Automatic Quality Assessment of Peptide Tandem Mass Spectra. Bioinformatics 20, i49–i54. https://doi.org/10.1093/bioinformatics/bth947

Chen, X., Jeong, J.C., 2007. Enhanced Recursive Feature Elimination, in: Proceedings of the Sixth International Conference on Machine Learning and Applications, ICMLA '07. IEEE Computer Society, USA, pp. 429–435. https://doi.org/10.1109/ICMLA.2007.44

Choo, K.W., Tham, W.M., 2007. Tandem mass spectrometry data quality assessment by self-convolution. BMC Bioinformatics 8, 352. https://doi.org/10.1186/1471-2105-8-352

Codrean, S., 2022. HRMS-Quality-assessment [WWW Document]. URL https://github.com/svetlanacodrean/HRMS-Quality-assessment (accessed 12.19.22).

Ding, J., Shi, J., Wu, F.-X., 2009. Quality Assessment of Tandem Mass Spectra by Using a Weighted K-Means. Clin. Proteomics 5, 15–22. https://doi.org/10.1007/s12014-009-9025-4

Dodder, N., Mullen, K., 2017. OrgMassSpecR: Organic Mass Spectrometry.

Dührkop, K., Shen, H., Meusel, M., Rousu, J., Böcker, S., 2015. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc. Natl. Acad. Sci. 112, 12580–12585. https://doi.org/10.1073/pnas.1509788112

Fiehnlab, 2022. MassBank of North America [WWW Document]. URL https://mona.fiehnlab.ucdavis.edu/ (accessed 11.22.22).

Gholamizoj, S., Ma, B., 2022. SPEQ: quality assessment of peptide tandem mass spectra with deep learning. Bioinformatics 38, 1568–1574. https://doi.org/10.1093/bioinformatics/btab874

Goutte, C., Gaussier, E., 2005. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation, in: Losada, D.E., Fernández-Luna, J.M. (Eds.), Advances in Information Retrieval, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 345–359. https://doi.org/10.1007/978-3-540-31865-1_25

Ham, J., Chen, Y., Crawford, M.M., Ghosh, J., 2005. Investigation of the random forest framework for classification of hyperspectral data. IEEE Trans. Geosci. Remote Sens. 43, 492–501. https://doi.org/10.1109/TGRS.2004.842481

Hollender, J., Schymanski, E.L., Singer, H.P., Ferguson, P.L., 2017. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? Environ. Sci. Technol. 51, 11505–11512. https://doi.org/10.1021/acs.est.7b02184

Liu, Y., Wang, Y., Zhang, J., 2012. New Machine Learning Algorithm: Random Forest, in: Liu, B., Ma, M., Chang, J. (Eds.), Information Computing and Applications, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 246–252. https://doi.org/10.1007/978-3-642-34062-8_32

Logan, B., Kontothanassis, L., Goddeau, D., Moreno, P.J., Hookway, R., Sarracino, D., 2004. Reducing the cost of protein identifications from mass spectrometry databases, in: The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 3060–3063. https://doi.org/10.1109/IEMBS.2004.1403865

Ma, C., 2017. DeepQuality: Mass Spectra Quality Assessment via Compressed Sensing and Deep Learning. ArXiv171011430 Q-Bio.

Mohammed Taha, H., Aalizadeh, R., Alygizakis, N., Antignac, J.-P., Arp, H.P.H., Bade, R., Baker, N., Belova, L., Bijlsma, L., Bolton, E.E., Brack, W., Celma, A., Chen, W.-L., Cheng, T., Chirsir, P., Čirka, Ľ., D'Agostino, L.A., Djoumbou Feunang, Y., Dulio, V., Fischer, S., Gago-Ferrero, P., Galani, A., Geueke, B., Głowacka, N., Glüge, J., Groh, K., Grosse, S., Haglund, P., Hakkinen, P.J., Hale, S.E., Hernandez, F., Janssen, E.M.-L., Jonkers, T., Kiefer, K., Kirchner, M., Koschorreck, J., Krauss, M., Krier, J., Lamoree, M.H., Letzel, M., Letzel, T., Li, Q., Little, J., Liu, Y., Lunderberg, D.M., Martin, J.W., McEachran, A.D., McLean, J.A., Meier, C., Meijer, J., Menger, F., Merino, C., Muncke, J., Muschket, M., Neumann, M., Neveu, V., Ng, K., Oberacher, H., O'Brien, J., Oswald, P., Oswaldova, M., Picache, J.A., Postigo, C., Ramirez, N., Reemtsma, T., Renaud, J., Rostkowski, P., Rüdel, H., Salek, R.M., Samanipour, S., Scheringer, M., Schliebner, I., Schulz, W., Schulze, T., Sengl, M., Shoemaker, B.A., Sims, K., Singer, H., Singh, R.R., Sumarah, M., Thiessen, P.A., Thomas, K.V., Torres, S., Trier, X., van Wezel, A.P., Vermeulen, R.C.H., Vlaanderen, J.J., von der Ohe, P.C., Wang, Z., Williams, A.J., Willighagen, E.L., Wishart, D.S., Zhang, J., Thomaidis, N.S., Hollender, J., Slobodnik, J., Schymanski, E.L., 2022. The NORMAN Suspect List Exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry. Environ. Sci. Eur. 34, 104. https://doi.org/10.1186/s12302-022-00680-6

Nesvizhskii, A.I., Roos, F.F., Grossmann, J., Vogelzang, M., Eddes, J.S., Gruissem, W., Baginsky, S., Aebersold, R., 2006. Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data: Toward More Efficient Identification of Post-translational Modifications, Sequence Polymorphisms, and Novel Peptides*. Mol. Cell. Proteomics 5, 652–670. https://doi.org/10.1074/mcp.M500319-MCP200

Neumann, S., Stravs, M., Schymanski, E., Schulze, T., 2022. MassBankEU [WWW Document]. URL https://massbank.eu/MassBank/Search (accessed 11.22.22).

Oberacher, H., Reinstadler, V., Kreidl, M., Stravs, M.A., Hollender, J., Schymanski, E.L., 2019. Annotating Nontargeted LC-HRMS/MS Data with Two Complementary Tandem Mass Spectral Libraries. Metabolites 9, 3. https://doi.org/10.3390/metabo9010003

Oberacher, H., Sasse, M., Antignac, J.-P., Guitton, Y., Debrauwer, L., Jamin, E.L., Schulze, T., Krauss, M., Covaci, A., Caballero-Casero, N., Rousseau, K., Damont, A., Fenaille, F., Lamoree, M., Schymanski, E.L., 2020. A European proposal for quality control and quality assurance of tandem mass spectral libraries. Environ. Sci. Eur. 32, 43. https://doi.org/10.1186/s12302-020-00314-9

Oberleitner, D., Schmid, R., Schulz, W., Bergmann, A., Achten, C., 2021. Feature-based molecular networking for identification of organic micropollutants including metabolites by non-target analysis applied to riverbank filtration. Anal. Bioanal. Chem. 413, 5291–5300. https://doi.org/10.1007/s00216-021-03500-7

Peets, P., Wang, W.-C., MacLeod, M., Breitholtz, M., Martin, J.W., Kruve, A., 2022. MS2Tox Machine Learning Tool for Predicting the Ecotoxicity of Unidentified Chemicals in Water by Nontarget LC-HRMS. Environ. Sci. Technol. 56, 15508–15517. https://doi.org/10.1021/acs.est.2c02536

Purushotham, S., Tripathy, B.K., 2012. Evaluation of Classifier Models Using Stratified Tenfold Cross Validation Techniques, in: Krishna, P.V., Babu, M.R., Ariwa, E. (Eds.), Global Trends in Information Systems and Software Applications, Communications in Computer and Information Science. Springer, Berlin, Heidelberg, pp. 680–690. https://doi.org/10.1007/978-3-642-29216-3_74

Schulze, B., Jeon, Y., Kaserzon, S., Heffernan, A.L., Dewapriya, P., O'Brien, J., Gomez Ramos, M.J., Ghorbani Gorji, S., Mueller, J.F., Thomas, K.V., Samanipour, S., 2020. An assessment of quality assurance/quality control efforts in high resolution mass spectrometry non-target workflows for analysis of environmental samples. TrAC Trends Anal. Chem. 133, 116063. https://doi.org/10.1016/j.trac.2020.116063

Schulze, T., Meier, R., Alygizakis, N., Schymanski, E., Bach, E., LI, D.H., lauperbe, raalizadeh, Tanaka, S., Witting, M., 2021. MassBank/MassBank-data: Release version 2021.12. https://doi.org/10.5281/zenodo.5775684

Stravs, M.A., Schymanski, E.L., Singer, H.P., Hollender, J., 2013. Automatic recalibration and processing of tandem mass spectra using formula annotation. J. Mass Spectrom. 48, 89–99. https://doi.org/10.1002/jms.3131

Tabb, D.L., Eng, J.K., Yates, J.R., 2001. Protein Identification by SEQUEST, in: Proteome Research: Mass Spectrometry. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 125–142. https://doi.org/10.1007/978-3-642-56895-4_7

Watrous, J., Roach, P., Alexandrov, T., Heath, B.S., Yang, J.Y., Kersten, R.D., van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J.M., Moore, B.S., Laskin, J., Bandeira, N., Dorrestein, P.C., 2012. Mass spectral molecular networking of living microbial colonies. Proc. Natl. Acad. Sci. 109, E1743–E1752. https://doi.org/10.1073/pnas.1203689109

Zou, A.-M., Wu, F.-X., Ding, J.-R., Poirier, G.G., 2009. Quality assessment of tandem mass spectra using support vector machine (SVM). BMC Bioinformatics 10, S49. https://doi.org/10.1186/1471-2105-10-S1-S49